
Efficient Algorithms and Error Analysis for the Modified Nyström Method

Shusen Wang

College of Computer Science & Technology
Zhejiang University, Hangzhou, China
wss@zju.edu.cn

Zhihua Zhang

Department of Computer Science & Engineering
Shanghai Jiao Tong University, Shanghai, China
zhizhua@sjtu.edu.cn

Abstract

Many kernel methods suffer from high time and space complexities and are thus prohibitive in big-data applications. To tackle the computational challenge, the Nyström method has been extensively used to reduce time and space complexities by sacrificing some accuracy. The Nyström method speeds up computation by constructing an approximation of the kernel matrix using only a few columns of the matrix. Recently, a variant of the Nyström method called the modified Nyström method has demonstrated significant improvement over the standard Nyström method in approximation accuracy, both theoretically and empirically. In this paper, we propose two algorithms that make the modified Nyström method practical. First, we devise a simple column selection algorithm with a provable error bound. Our algorithm is more efficient and easier to implement than and nearly as accurate as the state-of-the-art algorithm. Second, with the selected columns at hand, we propose an algorithm that computes the approximation in lower time complexity than the approach in the previous work. Furthermore, we prove that the modified Nyström method is exact under certain conditions, and we establish a lower error bound for the modified Nyström method.

1 Introduction

The kernel method is an important tool in machine learning, computer vision, and data mining (Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004).

However, many kernel methods require matrix computations of high time and space complexities. For example, let m be the number of data instances. The Gaussian process regression computes the inverse of an $m \times m$ matrix which takes time $\mathcal{O}(m^3)$ and space $\mathcal{O}(m^2)$; the kernel PCA, Isomap, and Laplacian eigenmaps all perform the truncated singular value decomposition which takes time $\mathcal{O}(m^2k)$ and space $\mathcal{O}(m^2)$, where k is the target rank of the decomposition. When m is large, it is challenging to store the $m \times m$ kernel matrix in RAM to perform these matrix computations. Therefore, these kernel methods are prohibitive when m is large.

To overcome the computational challenge, Williams and Seeger (2001) employed the Nyström method (Nyström, 1930) to generate a low-rank approximation to the original symmetric positive semidefinite (SPSD) kernel matrix. By using the Nyström method, eigenvalue decomposition and some matrix inverse can be approximately done on only a few columns of the SPSP matrix instead of on the entire matrix, and the time and space costs are reduced to $\mathcal{O}(m)$. The Nyström method has been widely used to speed up various kernel methods, such as the Gaussian process regression (Williams and Seeger, 2001), spectral clustering (Fowlkes et al., 2004; Li et al., 2011), kernel SVMs (Zhang et al., 2008; Yang et al., 2012), kernel PCA (Zhang et al., 2008; Zhang and Kwok, 2010; Talwalkar et al., 2013), kernel ridge regression (Cortes et al., 2010; Yang et al., 2012), determinantal processes (Affandi et al., 2013), etc.

To construct a low-rank matrix approximation, the Nyström method requires a small number of columns (say, c columns) to be selected from the kernel matrix by a column sampling technique. The approximation accuracy is largely determined by the sampling technique; that is, a better sampling technique can result in a Nyström approximate with a lower approximation error. In the previous work much attention has been made on improving the error bounds of the Nyström method: additive-error bound has been explored by Drineas and Mahoney (2005); Shawe-taylor et al. (2005); Kumar et al. (2012); Jin et al. (2012), etc. Very recently, Gittens and Mahoney (2013) established

Appearing in Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

the first relative-error bound which is more interesting than additive-error bound (Mahoney, 2011).

However, the approximation quality cannot be arbitrarily improved by devising a very good sampling technique. As shown theoretically by Wang and Zhang (2013), no matter what sampling technique is used to construct the Nyström approximation, the incurred error (in the spectral norm or the squared Frobenius norm) must grow with matrix size m at least linearly. Thus, the Nyström approximation can be very rough when m is large, unless large number columns are selected. As was pointed out by Cortes et al. (2010), the tighter kernel approximation leads to the better learning accuracy, so it is useful to find a kernel approximation model that is more accurate than the Nyström method.

To improve the approximation accuracy, Wang and Zhang (2013) proposed a new alternative called *the modified Nyström method* and a sampling algorithm for the modified Nyström method. The modified Nyström method can be applied in the same way exactly as the standard Nyström method to speedup kernel methods. The modified Nyström method has an advantage that the error does not grow with matrix size m . Therefore, by using the modified Nyström method instead of the standard Nyström method, a significantly smaller number of columns is needed to attain the same accuracy as the standard Nyström method.

However, it is much more expensive to construct the modified Nyström approximation than to construct the standard standard Nyström approximation. Furthermore, an efficient implementation of the modified Nyström method keeps till open. In this paper we seek to make the modified Nyström method efficient and practical.

Additionally, Kumar et al. (2009); Talwalkar and Ros-tamizadeh (2010) showed that the standard Nyström approximation is exact when the original kernel matrix is low-rank. Wang and Zhang (2013) proved the lower error bounds of the standard Nyström method. It is still open whether the modified Nyström method has similar properties. So we explore the theoretical properties of the modified Nyström method in this paper.

In sum, this paper offers the following contributions:

- We devise a column selection algorithm with provable error bound for the modified Nyström method. We call it *the uniform+adaptive² algorithm*. It is more efficient and much easier to implement than the near-optimal+adaptive algorithm of Wang and Zhang (2013), yet its error bound is comparable with the near-optimal+adaptive algorithm.
- We provide an efficient algorithm for computing the intersection matrix of the modified Nyström method. This algorithm can significantly reduce the time cost, especially when the kernel matrix is sparse.
- We show that the modified Nyström approximation

exactly recovers the original matrix under some conditions.

- We established a lower error bound for the modified Nyström method. We conjecture that the lower error bound is tight.

The remainder of this paper is organized as follows. In Section 2 we define the notation used in this paper. In Section 3 we formally define the Nyström approximation methods and introduce some column sampling algorithms. In Section 4 we present an efficient column sampling algorithm and its error analysis. In Section 5 we devise an algorithm that computes the modified Nyström approximation more efficiently. In Section 6 we empirically evaluate our proposed two algorithms. In Section 7 we explore some theoretical properties of the modified Nyström method.

2 Notation

The notation used in this paper follows that of Wang and Zhang (2013). For an $m \times n$ matrix $\mathbf{A} = [a_{ij}]$, we let $\mathbf{a}^{(i)}$ be its i -th row, \mathbf{a}_j be its j -th column, $\|\mathbf{A}\|_F = (\sum_{i,j} a_{ij}^2)^{1/2}$ be its Frobenius norm, and $\|\mathbf{A}\|_2 = \max_{\mathbf{x} \neq \mathbf{0}} \|\mathbf{A}\mathbf{x}\|_2 / \|\mathbf{x}\|_2$ be its spectral norm.

Letting $\rho = \text{rank}(\mathbf{A})$, we write the condensed singular value decomposition (SVD) of \mathbf{A} as $\mathbf{A} = \mathbf{U}_\mathbf{A} \Sigma_\mathbf{A} \mathbf{V}_\mathbf{A}^T$, where the (i, i) -th entry of $\Sigma_\mathbf{A} \in \mathbb{R}^{\rho \times \rho}$ is the i -th largest singular value of \mathbf{A} . We also let $\mathbf{U}_{\mathbf{A},k}$ and $\mathbf{V}_{\mathbf{A},k}$ be the first k ($< \rho$) columns of $\mathbf{U}_\mathbf{A}$ and $\mathbf{V}_\mathbf{A}$, respectively, and $\Sigma_{\mathbf{A},k}$ be the $k \times k$ top sub-block of $\Sigma_\mathbf{A}$. Then the $m \times n$ matrix $\mathbf{A}_k = \mathbf{U}_{\mathbf{A},k} \Sigma_{\mathbf{A},k} \mathbf{V}_{\mathbf{A},k}^T$ is the “closest” rank- k approximation to \mathbf{A} .

Based on SVD, the *matrix coherence* of the columns of \mathbf{A} relative to the best rank- k approximation to \mathbf{A} is defined by $\mu_k = \frac{n}{k} \max_j \|\mathbf{V}_{\mathbf{A},k}^{(j)}\|_2^2$. Let $\mathbf{A}^\dagger = \mathbf{V}_\mathbf{A} \Sigma_\mathbf{A}^{-1} \mathbf{U}_\mathbf{A}^T$ be the *Moore-Penrose inverse* of \mathbf{A} . When \mathbf{A} is nonsingular, the Moore-Penrose inverse is identical to the matrix inverse. Given another $m \times c$ matrix \mathbf{C} , we define $\mathcal{P}_\mathbf{C} \mathbf{A} = \mathbf{C} \mathbf{C}^\dagger \mathbf{A}$ as the projection of \mathbf{A} onto the column space of \mathbf{C} and $\mathcal{P}_{\mathbf{C},k} \mathbf{A} = \mathbf{C} \cdot \text{argmin}_{\text{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{C}\mathbf{X}\|_F$ as the rank restricted projection. It is obvious that $\|\mathbf{A} - \mathcal{P}_\mathbf{C} \mathbf{A}\|_F \leq \|\mathbf{A} - \mathcal{P}_{\mathbf{C},k} \mathbf{A}\|_F$.

Finally, we discuss the time complexities of the matrix operations mentioned above. For an $m \times n$ general matrix \mathbf{A} (assume $m \geq n$), it takes $\mathcal{O}(mn^2)$ flops to compute the full SVD and $\mathcal{O}(mnk)$ flops to compute the truncated SVD of rank k ($< n$). The computation of \mathbf{A}^\dagger takes $\mathcal{O}(mn^2)$ flops. It is worth mentioning that although multiplying an $m \times n$ matrix by an $n \times p$ matrix takes mnp flops, it can be performed in full parallel by partitioning the matrices into blocks. Thus, the time and space expense of large-scale matrix multiplication is not a challenge in real-world applications. We denote the time complexity of such a matrix multiplication by $T_{\text{Multiply}}(mnp)$, which

can be tremendously smaller than $\mathcal{O}(mnp)$ in parallel computing environment (Halko et al., 2011). An algorithm can still be efficient even if it demands large-scale matrix multiplications.

3 Previous Work

In Section 3.1 we introduce the standard and modified Nyström methods and discuss their advantages and disadvantages. In Section 3.2 we describe some commonly used column sampling algorithms.

3.1 The Nyström Methods

Given an $m \times m$ symmetric matrix \mathbf{A} , one needs to select $c (\ll m)$ columns of \mathbf{A} to form a matrix $\mathbf{C} \in \mathbb{R}^{m \times c}$ to construct the standard or modified Nyström approximation. Without loss of generality, \mathbf{A} and \mathbf{C} can be permuted such that

$$\mathbf{A} = \begin{bmatrix} \mathbf{W} & \mathbf{A}_{21}^T \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{A}_{21} \end{bmatrix}, \quad (1)$$

where \mathbf{W} is of size $c \times c$. The standard Nyström approximation is defined by

$$\tilde{\mathbf{A}}_c^{\text{nys}} \triangleq \mathbf{C} \mathbf{U}^{\text{nys}} \mathbf{C}^T = \mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T,$$

and the modified Nyström approximation is

$$\tilde{\mathbf{A}}_c^{\text{mod}} \triangleq \mathbf{C} \mathbf{U}^{\text{mod}} \mathbf{C}^T = \mathbf{C} (\mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T) \mathbf{C}^T.$$

Here the $c \times c$ matrices $\mathbf{U}^{\text{nys}} \triangleq \mathbf{W}^\dagger$ and $\mathbf{U}^{\text{mod}} \triangleq \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$ are called *the intersection matrices*. We see that the only difference between the two models is their intersection matrices.

For the approximation $\mathbf{C} \mathbf{U} \mathbf{C}^T$ constructed by either of the methods, given a target rank k , we hope the error ratio

$$f = \|\mathbf{A} - \mathbf{C} \mathbf{U} \mathbf{C}^T\|_\xi / \|\mathbf{A} - \mathbf{A}_k\|_\xi, \quad (\xi = F \text{ or } 2),$$

is as small as possible. However, Wang and Zhang (2013) showed that for the standard Nyström method, whatever a column selection algorithm is used, the ratio f must grow with the matrix size m when c is fixed.

Lemma 1 (Lower Error Bound of the Standard Nyström Method (Wang and Zhang, 2013)). *Whatever a column sampling algorithm is used, there exists an $m \times m$ SPSP matrix \mathbf{A} such that the error incurred by the standard Nyström method obeys:*

$$\begin{aligned} \|\mathbf{A} - \mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T\|_F^2 &\geq \Omega\left(1 + \frac{mk}{c^2}\right) \|\mathbf{A} - \mathbf{A}_k\|_F^2, \\ \|\mathbf{A} - \mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T\|_2 &\geq \Omega\left(\frac{m}{c}\right) \|\mathbf{A} - \mathbf{A}_k\|_2. \end{aligned}$$

Here k is an arbitrary target rank, and c is the number of selected columns.

Thus, when the matrix size m is large, the standard Nyström approximation is very inaccurate unless a large number of columns are selected. By comparison, when using an algorithm in Wang and Zhang (2013) for the modified Nyström method, the error ratio f remains constant for a fixed c and a growing m . Therefore, the modified Nyström method is more accurate than the standard Nyström method.

However, the accuracy gained by the modified Nyström method is at the cost of higher time and space complexities. Computing the intersection matrix $\mathbf{U}^{\text{nys}} = \mathbf{W}^\dagger$ only takes time $\mathcal{O}(c^3)$ and space $\mathcal{O}(c^2)$, while computing $\mathbf{U}^{\text{mod}} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$ naively takes time $\mathcal{O}(mc^2) + T_{\text{Multiply}}(m^2c)$ and space $\mathcal{O}(mc)$ ¹.

3.2 Sampling Algorithms for the Nyström Methods

The column selection problem has been widely studied in the theoretical computer science community (Boutsidis et al., 2011; Mahoney, 2011; Guruswami and Sinop, 2012) and the numerical linear algebra community (Gu and Eisenstat, 1996; Stewart, 1999), and numerous algorithms have been devised and analyzed. Here we focus on some theoretically guaranteed algorithms studied in the theoretical computer science community.

In the previous work much attention has been paid on improving column sampling algorithms such that the Nyström approximation is more accurate. Uniform sampling is the simplest and most time-efficient column selection algorithm, and it has provable error bounds when applied to the standard Nyström method (Gittens, 2011; Jin et al., 2012; Kumar et al., 2012; Gittens and Mahoney, 2013). To improve the approximation accuracy, many importance sampling algorithms have been proposed, among which the adaptive sampling of Deshpande et al. (2006) (see Algorithm 2) and the leverage score based sampling of Drineas et al. (2008); Ma et al. (2014) are widely studied. The leverage score based sampling has provable bounds when applied to the standard Nyström method (Gittens and Mahoney, 2013), and the adaptive sampling has provable bounds when applied to the modified Nyström method (Wang and Zhang, 2013). Besides, quadratic Rényi entropy based active subset selection (De Brabanter et al., 2010) and k -means clustering based selection (Zhang and Kwok, 2010) are also effective algorithms, but they do not have additive-error or relative-error bound.

Particularly, Wang and Zhang (2013) proposed an algorithm for the modified Nyström method by combining the near-optimal column sampling algorithm (Boutsidis et al.,

¹The matrix multiplication can be done blockwisely, that is, loading two small blocks into RAM to perform multiplication at a time. So the space cost of the matrix multiplication is $\mathcal{O}(mc)$ rather than $\mathcal{O}(m^2)$ (Wang and Zhang, 2013).

Algorithm 1 The Uniform+Adaptive² Algorithm.

- 1: **Input:** an $m \times m$ symmetric matrix \mathbf{A} , target rank k , error parameter $\epsilon \in (0, 1]$, matrix coherence μ .
 - 2: **Uniform Sampling.** Uniformly sample

$$c_1 = 8.7\mu k \log(\sqrt{5}k)$$
 columns of \mathbf{A} without replacement to construct \mathbf{C}_1 ;
 - 3: **Adaptive Sampling.** Sample

$$c_2 = 10k\epsilon^{-1}$$
 columns of \mathbf{A} to construct \mathbf{C}_2 using adaptive sampling algorithm 2 according to the residual $\mathbf{A} - \mathcal{P}_{\mathbf{C}_1} \mathbf{A}$;
 - 4: **Adaptive Sampling.** Sample

$$c_3 = 2\epsilon^{-1}(c_1 + c_2)$$
 columns of \mathbf{A} to construct \mathbf{C}_3 using adaptive sampling algorithm 2 according to the residual $\mathbf{A} - \mathcal{P}_{[\mathbf{C}_1, \mathbf{C}_2]} \mathbf{A}$;
 - 5: **return** $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3]$ and $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$.
-

2011) and the adaptive sampling algorithm (Deshpande et al., 2006). The error bound of the algorithm is the strongest among all the feasible algorithms for the Nyström methods. We show it in the following lemma.

Lemma 2 (The Near-Optimal+Adaptive Algorithm (Wang and Zhang, 2013)). *Given a symmetric matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ and a target rank k , the algorithm samples totally $c = \mathcal{O}(k\epsilon^{-2})$ columns of \mathbf{A} to construct the approximation. We run the algorithm $t \geq (2\epsilon^{-1} + 1) \log(1/p)$ times (independently in parallel) and choose the sample that minimizes $\|\mathbf{A} - \mathbf{C}(\mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T) \mathbf{C}^T\|_F$, then the inequality*

$$\|\mathbf{A} - \mathbf{C}(\mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T) \mathbf{C}^T\|_F \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F$$

holds with probability at least $1 - p$. The algorithm costs $\mathcal{O}(mc^2 + mk^3\epsilon^{-2/3}) + T_{\text{Multiply}}(m^2c)$ time and $\mathcal{O}(mc)$ space in computing \mathbf{C} and \mathbf{U} .

The near-optimal+adaptive algorithm is effective and efficient, but its implementation is very complicated. Its main component—the near-optimal column selection algorithm—consists of three steps: approximate SVD via random projection (Boutsidis et al., 2011; Halko et al., 2011), the dual-set sparsification algorithm (Boutsidis et al., 2011), and the adaptive sampling algorithm (Deshpande et al., 2006). Without careful implementation of the first two steps, the time and space costs roar, making the near-optimal+adaptive algorithm inefficient.

4 An Efficient Column Sampling Algorithm for the Modified Nyström Method

In this paper we propose a column sampling algorithm which is efficient, effective, and very easy to implement. The algorithm consists of a uniform sampling step and two adaptive sampling steps, so we call it the *uniform+adaptive² algorithm*. The algorithm is described in Algorithm 1 and analyzed in Theorem 3.

Algorithm 2 The Adaptive Sampling Algorithm.

- 1: **Input:** a residual matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$ and number of selected columns $c (< n)$.
 - 2: Compute sampling probabilities $p_j = \|\mathbf{b}_j\|_2^2 / \|\mathbf{B}\|_F^2$ for $j = 1, \dots, n$;
 - 3: Select c indices in c i.i.d. trials, in each trial the index j is chosen with probability p_j ;
 - 4: **return** an index set containing the indices of the selected columns.
-

The idea behind the uniform+adaptive² algorithm is quite intuitive. Since the modified Nyström method is the simultaneous projection of \mathbf{A} onto the column space of \mathbf{C} and the row space of \mathbf{C}^T , the approximation error will get lower if $\text{span}(\mathbf{C})$ better approximates $\text{span}(\mathbf{A})$. After the initialization by uniform sampling, the columns of \mathbf{A} far from $\text{span}(\mathbf{C}_1)$ have large residuals and are thus likely to get chosen by the adaptive sampling. After two rounds of adaptive sampling, columns of \mathbf{A} are likely to be near $\text{span}(\mathbf{C})$.

It is worth mentioning that our uniform+adaptive² algorithm is similar to the adaptive-full algorithm of (Kumar et al., 2012, Figure 3). The adaptive-full algorithm consists of a random initialization followed by multiple adaptive sampling steps. Obviously, using multiple adaptive sampling steps can surely reduce the approximation error. However, the update of sampling probability in each step is expensive, so we choose to do only two steps. Importantly, the adaptive-full algorithm of (Kumar et al., 2012, Figure 3) is merely a heuristic scheme without theoretical guarantee, whereas our uniform+adaptive² algorithm has a strong error bound which is nearly as good as the state-of-the-art algorithm of Wang and Zhang (2013) (See Theorem 3).

Theorem 3 (The Uniform+Adaptive² Algorithm.). *Given an $m \times m$ symmetric matrix \mathbf{A} and a target rank k , we let μ_k denote the matrix coherence of \mathbf{A} . Algorithm 1 samples totally*

$$c = \mathcal{O}(k\epsilon^{-2} + \mu_k\epsilon^{-1}k \log k)$$

columns of \mathbf{A} to construct the approximation. We run Algorithm 1

$$t \geq (20\epsilon^{-1} + 18) \log(1/p)$$

times (independently in parallel) and choose the sample that minimizes $\|\mathbf{A} - \mathbf{C}(\mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T) \mathbf{C}^T\|_F$, then the inequality

$$\|\mathbf{A} - \mathbf{C}(\mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T) \mathbf{C}^T\|_F \leq (1 + \epsilon) \|\mathbf{A} - \mathbf{A}_k\|_F$$

holds with probability at least $1 - p$. The algorithm costs $\mathcal{O}(mc^2) + T_{\text{Multiply}}(m^2c)$ time and $\mathcal{O}(mc)$ space in computing \mathbf{C} and \mathbf{U} .

Remark 1. *Theoretically, Algorithm 1 requires to compute the matrix coherence of \mathbf{A} in order to determine c_1 , c_2 , and c_3 . However, computing the matrix coherence*

Table 1: Comparisons between the two sampling algorithms in time complexity, space complexity, the number of selected columns, and the hardness of implementation.

	Uniform+Adaptive ²	Near-Optimal+Adaptive
Time	$\mathcal{O}(mc^2) + T_{\text{Multiply}}(m^2c)$	$\mathcal{O}(mc^2 + mk^3\epsilon^{-2/3}) + T_{\text{Multiply}}(m^2c)$
Space	$\mathcal{O}(mc)$	$\mathcal{O}(mc)$
#columns	$\mathcal{O}(k\epsilon^{-2} + \mu_k\epsilon^{-1}k \log k)$	$\mathcal{O}(k\epsilon^{-2})$
Implement	Easy to implement	Hard to implement

takes time $\mathcal{O}(m^2k)$ and is thus impractical; even the fast approximation approach of Drineas et al. (2012) is not feasible here because \mathbf{A} is a square matrix. The use of the matrix coherence here is merely for theoretical analysis; setting the parameter μ in Algorithm 1 to be exactly the matrix coherence does not certainly result in the highest accuracy. According to our off-line experiments, the resulting approximation accuracy is not sensitive to the value of μ . So we strongly suggest the users to set μ in Algorithm 1 to be a constant rather than actually computing the matrix coherence.

Table 1 presents comparisons between the near-optimal+adaptive algorithm of Wang and Zhang (2013) and our uniform+adaptive² algorithm. The time complexity of our algorithm is lower than the near-optimal+adaptive algorithm, and the space complexities of the two algorithms are the same. To attain the same error bound, our algorithm needs to select $c = \mathcal{O}(k\epsilon^{-2} + \mu_k\epsilon^{-1}k \log k)$ columns, which is a little larger than that of the near-optimal+adaptive algorithm. When $\epsilon \rightarrow 0$, we have that $\mathcal{O}(k\epsilon^{-2} + \mu_k\epsilon^{-1}k \log k) = \mathcal{O}(k\epsilon^{-2})$. Therefore, the error bound of our algorithm is nearly as good as the near-optimal+adaptive algorithm because ϵ is usually set to be a very small value.

5 Fast Computation of the Intersection Matrix

Naively computing the intersection matrix $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$ takes time $\mathcal{O}(mc^2) + T_{\text{Multiply}}(m^2c)$, which is much more expensive than computing \mathbf{W}^\dagger for the standard Nyström method. In this section we propose a more efficient algorithm for computing the intersection matrix, which only takes time $\mathcal{O}(c^3) + T_{\text{Multiply}}((m-c)^2c)$. The algorithm is described in Theorem 4. The algorithm is obtained by expanding the Moore-Penrose inverse of \mathbf{C} using the theorem in (Ben-Israel and Greville, 2003, Page 179).

Theorem 4. For an $m \times m$ symmetric matrix \mathbf{A} , when the submatrix \mathbf{W} is nonsingular, the intersection matrix of the modified Nyström method $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$ can be computed in time $\mathcal{O}(c^3) + T_{\text{Multiply}}((m-c)^2c)$ by the

Table 2: A summary of the datasets for the Nyström approximation.

Dataset	#Instance	#Attribute	Source
Letters	15,000	16	Michie et al. (1994)
Abalone	4,177	8	Frank and Asuncion (2010)
Wine Quality	4,898	12	Cortez et al. (2009)

following formula:

$$\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T = \mathbf{T}_1 (\mathbf{W} + \mathbf{T}_2 + \mathbf{T}_2^T + \mathbf{T}_3) \mathbf{T}_1^T,$$

where the intermediate matrices are computed by

$$\begin{aligned} \mathbf{T}_0 &= \mathbf{A}_{21}^T \mathbf{A}_{21}, & \mathbf{T}_1 &= \mathbf{W}^{-1} (\mathbf{I}_c + \mathbf{W}^{-1} \mathbf{T}_2)^{-1}, \\ \mathbf{T}_2 &= \mathbf{T}_0 \mathbf{W}^{-1}, & \mathbf{T}_3 &= \mathbf{W}^{-1} (\mathbf{A}_{21}^T \mathbf{A}_{22} \mathbf{A}_{21}) \mathbf{W}^{-1}. \end{aligned}$$

The four intermediate matrices are all of size $c \times c$, and the matrix inverse operations are on $c \times c$ small matrices.

Remark 2. Since the submatrix \mathbf{W} is not in general nonsingular, before using the algorithm, the user should first test the rank of \mathbf{W} , which takes time $\mathcal{O}(c^3)$. Empirically, for graph Laplacian and the radial basis function (RBF) kernel (Genton, 2001), the submatrix \mathbf{W} is usually nonsingular, and the algorithm is useful; for the linear kernel, \mathbf{W} is often singular, so the algorithm does not work.

6 Experiments

In this section we empirically evaluate our two algorithms proposed in Section 4 and 5. In Section 6.1 we compare the sampling algorithms for the modified Nyström method in terms of approximation error and time expense. In Section 6.2 we illustrate the effect of our algorithm for computing the intersection matrix $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$.

We implement all of the compared algorithms in MATLAB and conduct experiments on a workstation with Intel Xeon 2.40GHz CPUs, 24GB RAM, and 64bit Windows Server 2008 system. To compare the running time, all the computations are carried out in a single thread in MATLAB.

6.1 Comparisons among the Sampling Algorithms

We mainly compare our uniform+adaptive² algorithm (Algorithm 1) with the near-optimal+adaptive algorithm (Wang and Zhang, 2013); the two algorithms are the only provable algorithms for the modified Nyström method. We also employ the uniform sampling and the leverage-score based sampling (Drineas et al., 2008; Gittens and Mahoney, 2013) as baselines (they are widely used but not provable for the modified Nyström method).

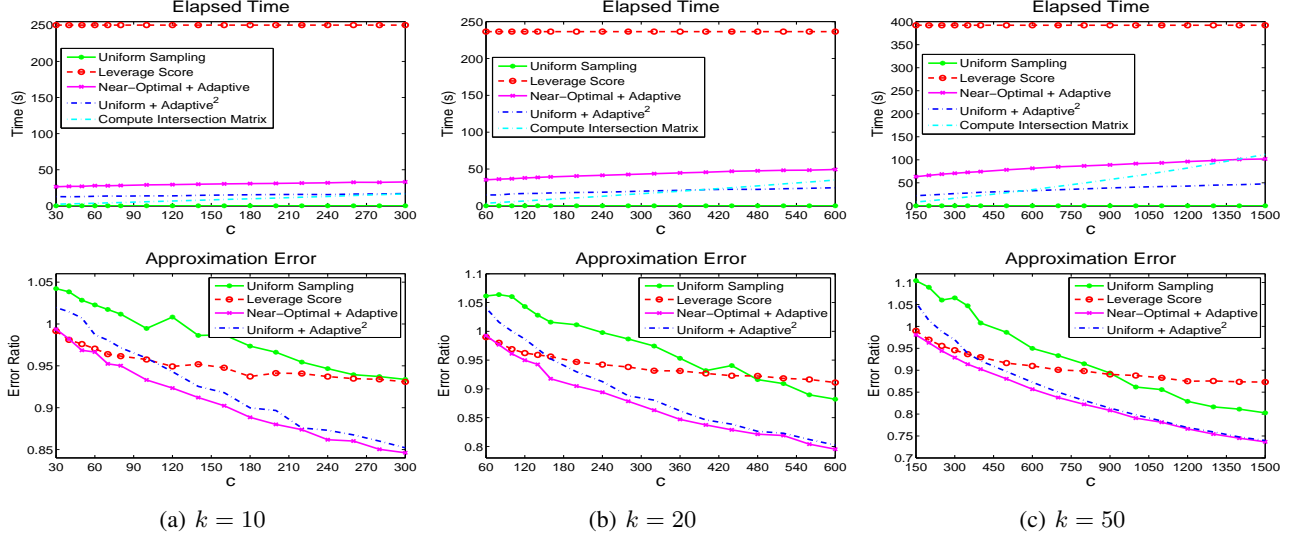


Figure 1: Results on the RBF kernel of the Letters dataset. Here the matrix coherence of the kernel matrix is $\mu_{10} = 62.05$, $\mu_{20} = 34.87$, and $\mu_{50} = 19.16$.

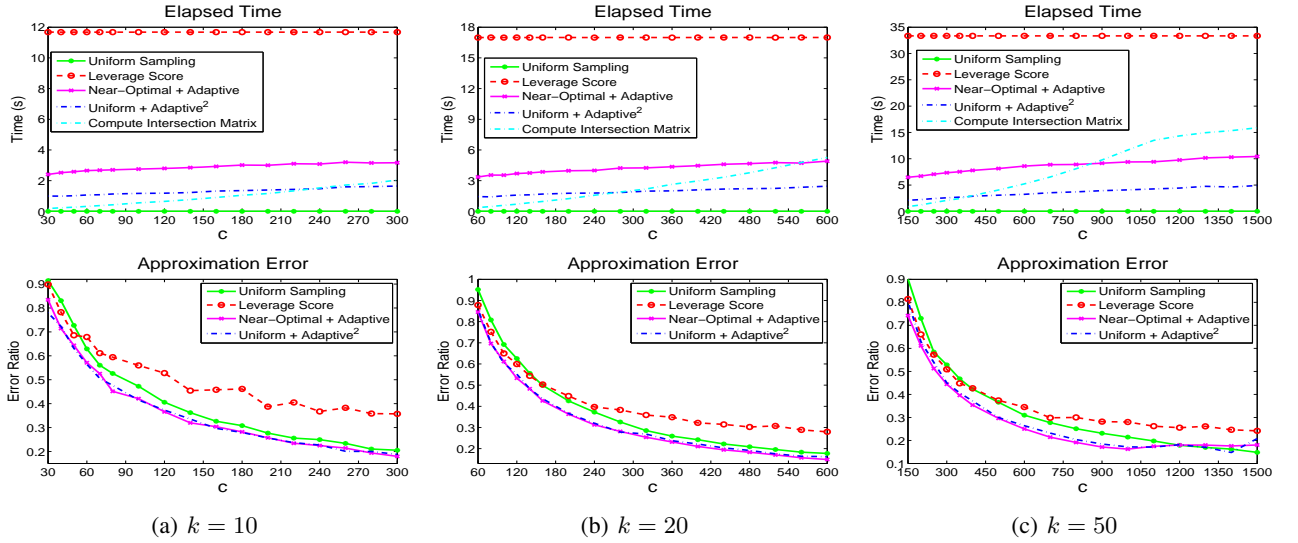


Figure 2: Results on the RBF kernel of the Abalone dataset. Here the matrix coherence of the kernel matrix is $\mu_{10} = 3.28$, $\mu_{20} = 3.02$, and $\mu_{50} = 2.64$.

For all of the four algorithms, columns are sampled without replacement.

The experiment settings follows Wang and Zhang (2013). We report the approximation error and running time of each algorithm on each dataset. The approximation error is defined by

$$\text{Approximation Error} = \frac{\|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{C}^T\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F},$$

where k is a fixed target rank and \mathbf{U} is the intersection matrix.

We test the algorithms on three datasets summarized in Table 2. For each dataset we generate an RBF kernel matrix \mathbf{A} with $a_{ij} = \exp(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$, where \mathbf{x}_i and \mathbf{x}_j are data instances and σ is the parameter defining the scale of the kernel. We set $\sigma = 0.2$ in our experiments. For each dataset we fix a target rank $k = 10, 20$, or 50 , and vary c in a very large range. We run each algorithm for 20 times and report the the minimum approximation error of the 20 repeats. We also report the average elapsed time of column selection and the computation of the $c \times c$ intersection matrix, respectively. Here we report the *average* elapsed time rather than the total time of the 20 repeats because the

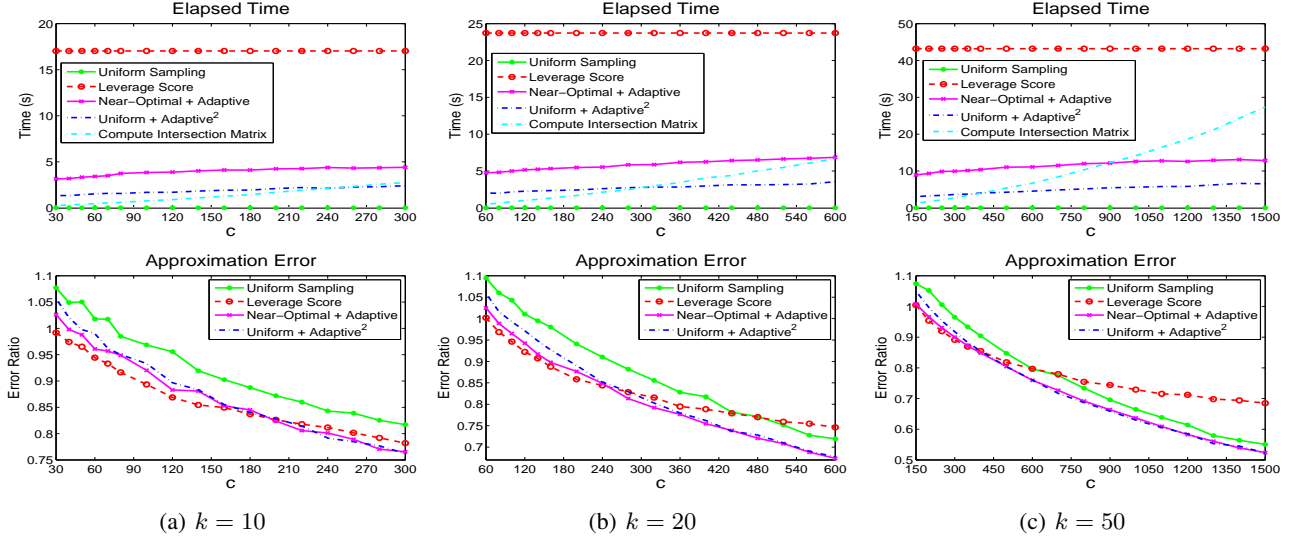


Figure 3: Results on the RBF kernel of the Wine Quality dataset. Here the matrix coherence of the kernel matrix is $\mu_{10} = 16.17$, $\mu_{20} = 12.13$, and $\mu_{50} = 9.30$.

20 repeats can be performed in parallel. The results are depicted in Figures 1, 2, and 3.

The empirical results in the figures show that our uniform+adaptive² algorithm achieves accuracy comparable with the state-of-the-art algorithm—the near-optimal+adaptive algorithm of Wang and Zhang (2013). Especially, when c is large, those two algorithms have virtually the same accuracy, which is in accordance with our analysis in the last paragraph of Section 4: large c implies small error term ϵ , and the error bounds of the two algorithms coincide when ϵ is small. We can also see that our uniform+adaptive² algorithm works nearly as good as the near-optimal+adaptive algorithm when the matrix coherence μ_k is small (e.g. Figure 2); when the matrix coherence is large (e.g. Figure 1), the error of our algorithm is a little worse than the near-optimal+adaptive algorithm. Furthermore, our uniform+adaptive² algorithm is much more accurate than uniform sampling and the leverage-score based sampling in most cases.

As for the running time, we can see that our algorithm performs column selection very efficiently and the elapsed time grows slowly in c . By comparison, our algorithm is much more efficient than the other two nonuniform sampling algorithms.

6.2 Effect of the Fast Computation of the Intersection Matrix

To illustrate the effect of our algorithm for computing the intersection matrix $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$, we generate a kernel matrix of the Letters Dataset (Michie et al., 1994) which has 15,000 instances and 16 attributes. We first generate a dense RBF kernel matrix with scale parameter $\sigma = 0.2$,

and then obtain a sparse symmetric matrix by truncating the entries with small magnitude such that 1% entries are nonzero. We illustrate in Figure 4 the speedup induced by our algorithm. In both cases, our algorithm is faster than the naive approach, and the speedup is particularly significant when \mathbf{A} is sparse.

7 Theoretical Analysis for the Modified Nyström Method

In Section 7.1 we show that the modified Nyström approximation is exact when \mathbf{A} is low-rank. In Section 7.2 we provide a lower error bound of the modified Nyström method.

7.1 Theoretical Justifications

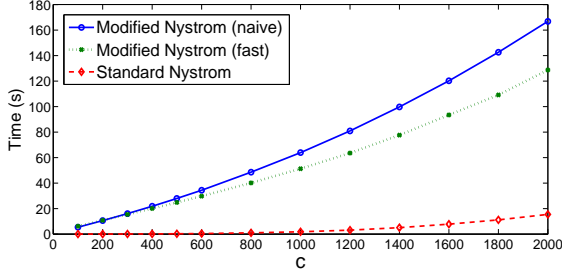
Kumar et al. (2009); Talwalkar and Rostamizadeh (2010) showed that the standard Nyström method is exact when $\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{A})$. We show in Theorem 5 a similar result for the modified Nyström approximations.

Theorem 5. For a symmetric matrix \mathbf{A} defined in (1), the following three statements are equivalent: (i) $\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{A})$, (ii) $\mathbf{A} = \mathbf{C}\mathbf{W}^\dagger \mathbf{C}^T$, (iii) $\mathbf{A} = \mathbf{C}\mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T \mathbf{C}^T$.

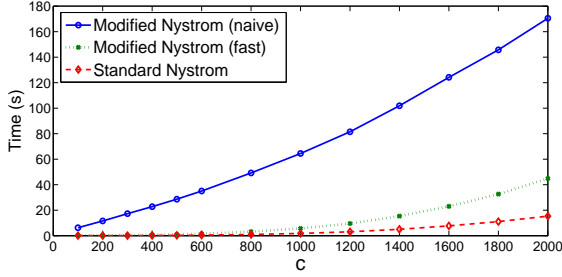
Theorem 5 shows that the standard and modified Nyström methods are equivalent when $\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{A})$. However, it holds in general that $\text{rank}(\mathbf{A}) \gg c \geq \text{rank}(\mathbf{W})$, where the two models are not equivalent.

Furthermore, $\mathbf{U}^{\text{mod}} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$ is the minimizer of the following minimization problem

$$\min_{\mathbf{U}} \|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{C}^T\|_F,$$



(a) Dense RBF kernel matrix.



(b) Sparse RBF kernel matrix with 1% nonzero entries.

Figure 4: Effect of our fast computation of the intersection matrix. The two matrices are both of size $15,000 \times 15,000$, and we sample c columns uniformly to compute the intersection matrix $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$ (the modified Nyström) and $\mathbf{U} = \mathbf{W}^\dagger$ (the standard Nyström). The time for computing \mathbf{U} is plotted in the figures.

so we have that

$$\|\mathbf{A} - \mathbf{C}(\mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T) \mathbf{C}\|_F \leq \|\mathbf{A} - \mathbf{C} \mathbf{W}^\dagger \mathbf{C}\|_F.$$

This shows that in general the modified Nyström method is more accurate than the standard Nyström method.

7.2 Lower Error Bound of the Modified Nyström Method

We establish in Theorem 6 a lower error bound of the modified Nyström method. Theorem 6 shows that whatever a column sampling algorithm is used to construct the modified Nyström approximation, at least $c \geq 2k\epsilon^{-1}$ columns must be chosen to attain the $1 + \epsilon$ bound.

Theorem 6 (Lower Error Bound of the Modified Nyström Method). *Whatever a column sampling algorithm is used, there exists an $m \times m$ SPSD matrix \mathbf{A} such that the error incurred by the modified Nyström method obeys:*

$$\|\mathbf{A} - \mathbf{C} \mathbf{U} \mathbf{C}^T\|_F^2 \geq \frac{m-c}{m-k} \left(1 + \frac{2k}{c}\right) \|\mathbf{A} - \mathbf{A}_k\|_F^2.$$

Here k is an arbitrary target rank, c is the number of selected columns, and $\mathbf{U} = \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T$.

Boutsidis et al. (2011) established a lower error bound for the column selection problem, and the lower error bound is

tight because it is attained by the optimal column selection algorithm of Guruswami and Sinop (2012). Boutsidis et al. (2011) showed that whatever column sampling algorithm is used, there exists an $m \times n$ matrix \mathbf{A} such that the error incurred by the projection of \mathbf{A} onto the column space of \mathbf{C} is lower bounded by

$$\|\mathbf{A} - \mathbf{C} \mathbf{C}^\dagger \mathbf{A}\|_F^2 \geq \frac{n-c}{n-k} \left(1 + \frac{k}{c}\right) \|\mathbf{A} - \mathbf{A}_k\|_F^2, (2)$$

where k is an arbitrary target rank, c is the number of selected columns.

Interestingly, the modified Nyström approximation is the projection of \mathbf{A} onto the column space of \mathbf{C} and the row space of \mathbf{C}^T simultaneously, so there is a strong resemblance between the modified Nyström approximation and the column selection problem. As we see, the lower error bound of the modified Nyström approximation in Theorem 6 differs from (2) only by a factor of 2. So it is a reasonable conjecture that the lower bound in Theorem 6 is tight, as well as the lower bound of the column selection problem in (2). We leave it as an open problem.

8 Conclusions and Future Work

In this paper we have proposed two algorithms to make the modified Nyström method more practical. First, we have proposed a column selection algorithm called *uniform+adaptive*² and provided an relative-error bound for the algorithm. The algorithm is highly efficient and effective and very easy to implement. The error bound of the algorithm is nearly as strong as that of the state-of-the-art algorithm—the near-optimal+adaptive algorithm—which is complicated. The experimental results have shown that our uniform+adaptive² algorithm is more efficient than the near-optimal+adaptive algorithm, while their accuracies are comparable. Second, we have devised an algorithm for computing the intersection matrix of the modified Nyström approximation; under certain conditions, our algorithm can significantly improve the time complexity. The speedup induced by this algorithm has also been verified empirically.

Furthermore, we have proved that the modified Nyström approximation can be exact when the original matrix is low-rank. We have also established a lower error bound for the modified Nyström method: at least $c \geq 2k\epsilon^{-1}$ columns must be chosen to attain the $1 + \epsilon$ bound. We have conjectured this lower error bound to be tight. Notice that the best known algorithm for the modified Nyström method requires at most $c = k\epsilon^{-2}$ columns to attain the $1 + \epsilon$ bound, so there is a gap between the lower and upper error bounds. It remains an open problem that if there exists an algorithm attaining the lower error bound.

Acknowledgement

This work has been supported in part by the Natural Science Foundation of China (No. 61070239), Microsoft Research Asia Fellowship 2013, and the Scholarship Award for Excellent Doctoral Student granted by Chinese Ministry of Education.

References

- Affandi, R. H., A. Kulesza, E. B. Fox, and B. Taskar (2013). Nyström approximation for large-scale determinantal processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Ben-Israel, A. and T. N. Greville (2003). *Generalized Inverses: Theory and Applications. Second Edition*. Springer.
- Boutsidis, C., P. Drineas, and M. Magdon-Ismael (2011). Near optimal column-based matrix reconstruction. In *Annual Symposium on Foundations of Computer Science (FOCS)*.
- Cortes, C., M. Mohri, and A. Talwalkar (2010). On the impact of kernel approximation on learning accuracy. In *Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Cortez, P., A. Cerdeira, F. Almeida, T. Matos, and J. Reis (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* 47(4), 547–553.
- De Brabanter, K., J. De Brabanter, J. A. Suykens, and B. De Moor (2010). Optimized fixed-size kernel models for large data sets. *Computational Statistics & Data Analysis* 54(6), 1484–1504.
- Deshpande, A., L. Rademacher, S. Vempala, and G. Wang (2006). Matrix approximation and projective clustering via volume sampling. *Theory of Computing* 2(2006), 225–247.
- Drineas, P., M. Magdon-Ismael, M. W. Mahoney, and D. P. Woodruff (2012). Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research* 13, 3441–3472.
- Drineas, P. and M. W. Mahoney (2005). On the Nyström method for approximating a gram matrix for improved kernel-based learning. *Journal of Machine Learning Research* 6, 2153–2175.
- Drineas, P., M. W. Mahoney, and S. Muthukrishnan (2008, September). Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications* 30(2), 844–881.
- Fowlkes, C., S. Belongie, F. Chung, and J. Malik (2004). Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(2), 214–225.
- Frank, A. and A. Asuncion (2010). UCI machine learning repository.
- Genton, M. G. (2001). Classes of kernels for machine learning: A statistics perspective. *Journal of Machine Learning Research* 2, 299–312.
- Gittens, A. (2011). The spectral norm error of the naive Nyström extension. *arXiv preprint arXiv:1110.5305*.
- Gittens, A. and M. W. Mahoney (2013). Revisiting the nyström method for improved large-scale machine learning. In *International Conference on Machine Learning (ICML)*.
- Gu, M. and S. C. Eisenstat (1996). Efficient algorithms for computing a strong rank-revealing QR factorization. *SIAM Journal on Scientific Computing* 17(4), 848–869.
- Guruswami, V. and A. K. Sinop (2012). Optimal column-based low-rank matrix reconstruction. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*.
- Halko, N., P.-G. Martinsson, and J. A. Tropp (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review* 53(2), 217–288.
- Jin, R., T. Yang, M. Mahdavi, Y.-F. Li, and Z.-H. Zhou (2012). Improved bounds for the Nyström method with application to kernel classification. *CoRR abs/1111.2262*.
- Kumar, S., M. Mohri, and A. Talwalkar (2009). On sampling-based approximate spectral decomposition. In *International Conference on Machine Learning (ICML)*.
- Kumar, S., M. Mohri, and A. Talwalkar (2012). Sampling methods for the Nyström method. *Journal of Machine Learning Research* 13, 981–1006.
- Li, M., X.-C. Lian, J. T. Kwok, and B.-L. Lu (2011). Time and space efficient spectral clustering via column sampling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ma, P., M. Mahoney, and B. Yu (2014). A statistical perspective on algorithmic leveraging. In *International Conference on Machine Learning (ICML)*.
- Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning* 3(2), 123–224.
- Michie, D., D. J. Spiegelhalter, and C. C. Taylor (1994). *Machine Learning, Neural and Statistical Classification*. Prentice Hall.
- Nyström, E. J. (1930). Über die praktische auflösung von integralgleichungen mit anwendungen auf randwertaufgaben. *Acta Mathematica* 54(1), 185–204.
- Schölkopf, B. and A. J. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.
- Shawe-Taylor, J. and N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press.
- Shawe-taylor, J., C. K. I. Williams, N. Cristianini, and J. Kandola (2005). On the eigenspectrum of the gram matrix and the generalisation error of kernel pca. *IEEE Transactions on Information Theory* 51, 2510–2522.
- Stewart, G. W. (1999). Four algorithms for the efficient computation of truncated pivoted QR approximations to a sparse matrix. *Numerische Mathematik* 83(2), 313–323.
- Talwalkar, A., S. Kumar, M. Mohri, and H. Rowley (2013). Large-scale svd and manifold learning. *Journal of Machine Learning Research* 14, 3129–3152.
- Talwalkar, A. and A. Rostamizadeh (2010). Matrix coherence and the Nyström method. *Conference on Uncertainty in Artificial Intelligence (UAI)*.
- Tropp, J. A. (2011). Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis* 3(01–02), 115–126.
- Wang, S. and Z. Zhang (2013). Improving CUR matrix decomposition and the Nyström approximation via adaptive sampling. *Journal of Machine Learning Research* 14, 2729–2769.
- Williams, C. and M. Seeger (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems (NIPS)*.

Yang, T., Y.-F. Li, M. Mahdavi, R. Jin, and Z.-H. Zhou (2012). Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems (NIPS)*.

Zhang, K. and J. T. Kwok (2010). Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks* 21(10), 1576–1587.

Zhang, K., I. W. Tsang, and J. T. Kwok (2008). Improved Nyström low-rank approximation and error analysis. In *International Conference on Machine Learning (ICML)*.

A Proof of Theorem 3

The error analysis for the uniform+adaptive² algorithm relies on Lemma 7, which guarantees the error incurred by its uniform sampling step. The proof of Lemma 7 essentially follows Gittens (2011). We prove Lemma 7 using probability inequalities and some techniques of Boutsidis et al. (2011); Gittens (2011); Gittens and Mahoney (2013); Tropp (2011); the proof is in Appendix A.1.

Lemma 7 (Uniform Column Sampling). *Given an $m \times n$ matrix \mathbf{A} and a target rank k , let μ_k denote the matrix coherence of \mathbf{A} . By sampling*

$$c = \frac{\mu_k k \log(k/\delta)}{\theta \log \theta - \theta + 1},$$

columns uniformly without replacement to construct \mathbf{C} , the following inequality

$$\|\mathbf{A} - \mathcal{P}_{\mathbf{C},k} \mathbf{A}\|_F^2 \leq (1 + \delta^{-1} \theta^{-1}) \|\mathbf{A} - \mathbf{A}_k\|_F^2.$$

holds with probability at least $1 - 2\delta$. Here $\delta \in (0, 0.5)$ and $\theta \in (0, 1)$ are arbitrary real numbers.

The error analysis for the two adaptive sampling steps of the uniform+adaptive² algorithm relies on Lemma 8, which follows immediately from (Wang and Zhang, 2013, Corollary 7 and Section 4.5).

Lemma 8. *Given an $m \times m$ symmetric matrix \mathbf{A} and a target rank k , we let \mathbf{C}_1 contain the c_1 columns of \mathbf{A} selected by a column sampling algorithm such that the following inequality holds:*

$$\|\mathbf{A} - \mathcal{P}_{\mathbf{C}_1} \mathbf{A}\|_F^2 \leq f \|\mathbf{A} - \mathbf{A}_k\|_F^2.$$

Then we select $c_2 = kf\epsilon^{-1}$ columns to construct \mathbf{C}_2 and $c_3 = (c_1 + c_2)\epsilon^{-1}$ columns to construct \mathbf{C}_3 , both using the adaptive sampling according to the residual $\mathbf{B}_1 = \mathbf{A} - \mathcal{P}_{\mathbf{C}_1} \mathbf{A}$ and $\mathbf{B}_2 = \mathbf{A} - \mathcal{P}_{[\mathbf{C}_1, \mathbf{C}_2]} \mathbf{A}$, respectively. Let $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3]$, we have that

$$\mathbb{P} \left\{ \frac{\|\mathbf{A} - \mathbf{C}(\mathbf{C}^\dagger \mathbf{A}(\mathbf{C}^\dagger)^T) \mathbf{C}^T\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F} \geq 1 + s\epsilon \right\} \leq \frac{1 + \epsilon}{1 + s\epsilon},$$

where s is an arbitrary constant greater than 1.

Finally Theorem 3 is proved by combining Lemma 7 and Lemma 8. The proof is in Appendix A.2.

A.1 Proof of Lemma 7

Proof. We use uniform column sampling to select c column of \mathbf{A} to construct $\mathbf{C} = \mathbf{A}\mathbf{S}$. Here the $n \times c$ random matrix \mathbf{S} has one entry equal to one and the rest equal to zero in each column, and at most one nonzero entry in each row, and \mathbf{S} is uniformly distributed among $\binom{n}{c}$ such kind of matrices. Applying Lemma 7 of Boutsidis et al. (2011), we get

$$\begin{aligned} \|\mathbf{A} - \mathcal{P}_{\mathbf{C},k} \mathbf{A}\|_F^2 &\leq \|\mathbf{A} - \mathbf{A}_k\|_F^2 + \|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\|_F^2 \|(\mathbf{V}_{\mathbf{A},k}^T \mathbf{S})^\dagger\|_2^2. \end{aligned} \quad (3)$$

Now we bound $\|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\|_F^2$ and $\|(\mathbf{V}_{\mathbf{A},k}^T \mathbf{S})^\dagger\|_2^2$ respectively using the techniques of Gittens (2011); Gittens and Mahoney (2013); Tropp (2011).

Let $\mathcal{I} \subset [n]$ be a random index set corresponding to \mathbf{S} . The support of \mathcal{I} is uniformly distributing among all the index sets in $2^{[n]}$ with cardinality c . According to Gittens and Mahoney (2013), the expectation of $\|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\|_F^2$ can be written as

$$\begin{aligned} \mathbb{E} \|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\|_F^2 &= \mathbb{E} \|(\mathbf{A} - \mathbf{A}_k)_{\mathcal{I}}\|_F^2 \\ &= c \mathbb{E} \|(\mathbf{A} - \mathbf{A}_k)_i\|_F^2 = \frac{c}{n} \|\mathbf{A} - \mathbf{A}_k\|_F^2. \end{aligned}$$

Applying Markov's inequality, we have that

$$\begin{aligned} \mathbb{P} \left\{ \|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\|_F^2 \geq \frac{c}{n\delta} \|\mathbf{A} - \mathbf{A}_k\|_F^2 \right\} &\leq \frac{\mathbb{E} \|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\|_F^2}{\frac{c}{n\delta} \|\mathbf{A} - \mathbf{A}_k\|_F^2} = \delta. \end{aligned} \quad (4)$$

Here $\delta \in (0, 0.5)$ is a real number defined later.

Now we establish the bound for $\mathbb{E} \|\Omega_2^\dagger\|_2^2$ as follows. Let $\lambda_i(\mathbf{X})$ be the i -th largest eigenvalue of \mathbf{X} . Following the proof of Lemma 1 of Gittens (2011), we have

$$\begin{aligned} \|(\mathbf{V}_{\mathbf{A},k}^T \mathbf{S})^\dagger\|_2^2 &= \lambda_k^{-1} \left(\mathbf{V}_{\mathbf{A},k}^T \mathbf{S} \mathbf{S}^T \mathbf{V}_{\mathbf{A},k} \right) \\ &= \lambda_k^{-1} \left(\sum_{i=1}^c \mathbf{X}_i \right) \leq \lambda_{\min}^{-1} \left(\sum_{i=1}^c \mathbf{X}_i \right), \end{aligned} \quad (5)$$

where the random matrices $\mathbf{X}_1, \dots, \mathbf{X}_c$ are chosen uniformly at random from the set $\left\{ (\mathbf{V}_{\mathbf{A},k}^T)_i (\mathbf{V}_{\mathbf{A},k}^T)_i^T \right\}_{i=1}^n$ without replacement. The random matrices are of size $k \times k$. We accordingly define

$$R = \max_i \lambda_{\max}(\mathbf{X}_i) = \max_i \|(\mathbf{V}_{\mathbf{A},k}^T)_i\|_2^2 = \frac{k}{n} \mu_k,$$

where μ_k is the matrix coherence of \mathbf{A} , and define

$$\begin{aligned} \beta_{\min} &= c \lambda_{\min}(\mathbb{E} \mathbf{X}_1) \\ &= \lambda_{\min} \left(\frac{c}{n} \mathbf{V}_{\mathbf{A},k}^T \mathbf{V}_{\mathbf{A},k} \right) = \frac{c}{n}. \end{aligned}$$

Then we apply Lemma 9 and obtained the following inequality:

$$\mathbb{P}\left[\lambda_{\min}\left(\sum_{i=1}^c \mathbf{X}_i\right) \leq \frac{\theta c}{n}\right] \leq k \left[\frac{e^{\theta-1}}{\theta^\theta}\right]^{\frac{c}{k\mu_k}} \triangleq \delta, \quad (6)$$

where $\theta \in (0, 1]$ is a real number, and it follows that

$$c = \frac{\mu_k k \log(k/\delta)}{\theta \log \theta - \theta + 1}.$$

Applying (5) and (6), we have

$$\mathbb{P}\left\{\|(\mathbf{V}_{\mathbf{A},k}^T \mathbf{S})^\dagger\|_2^2 \geq \frac{n}{\theta c}\right\} \leq \delta. \quad (7)$$

Combining (4) and (7) and applying the union bound, we have the following inequality:

$$\begin{aligned} \mathbb{P}\left\{\|(\mathbf{A} - \mathbf{A}_k)\mathbf{S}\|_F^2 \geq \frac{c}{n\delta}\|\mathbf{A} - \mathbf{A}_k\|_F^2 \right. \\ \left. \text{or } \|(\mathbf{V}_{\mathbf{A},k}^T \mathbf{S})^\dagger\|_2^2 \geq \frac{n}{\theta c}\right\} \leq 2\delta. \end{aligned} \quad (8)$$

Finally, from (3) and (8) we have that the inequality

$$\|\mathbf{A} - \mathcal{P}_{\mathbf{C},k}\mathbf{A}\|_F^2 \leq (1 + \delta^{-1}\theta^{-1})\|\mathbf{A} - \mathbf{A}_k\|_F^2$$

holds with probability at least $1 - 2\delta$, by which the lemma follows. \square

Lemma 9 (Theorem 2.2 of Tropp (2011)). *We are given l independent random $d \times d$ SPSP matrices $\mathbf{X}_1, \dots, \mathbf{X}_l$ with the property*

$$\lambda_{\max}(\mathbf{X}_i) \leq R \quad \text{for } i = 1, \dots, l.$$

We define $\mathbf{Y} = \sum_{i=1}^l \mathbf{X}_i$ and $\beta_{\min} = l\lambda_{\min}(\mathbb{E}\mathbf{X}_1)$. Then for any $\theta \in (0, 1]$, the following inequality holds:

$$\mathbb{P}\left\{\lambda_{\min}(\mathbf{Y}) \leq \theta \beta_{\min}\right\} \leq d \left[\frac{e^{\theta-1}}{\theta^\theta}\right]^{\frac{\beta_{\min}}{R}}.$$

A.2 Proof of the Theorem

Proof. The matrix \mathbf{C}_1 consists of c_1 columns selected by uniform sampling, and $\mathbf{C}_2 \in \mathbb{R}^{m \times c_2}$ and $\mathbf{C}_3 \in \mathbb{R}^{m \times c_3}$ are constructed by adaptive sampling. We set $\delta = 1/\sqrt{5}$ and $\theta = \sqrt{5}/4$ for Lemma 7, then we have

$$\begin{aligned} f &= 1 + \delta^{-1}\theta^{-1} = 5, \\ c_1 &= \frac{\mu_k k \log(k/\delta)}{\theta \log \theta - \theta + 1} = 8.7\mu_k k \log(\sqrt{5}k). \end{aligned}$$

Then we set

$$\begin{aligned} c_2 &= kf\epsilon^{-1} = 5k\epsilon^{-1}, \\ c_3 &= (c_1 + c_2)\epsilon^{-1}, \end{aligned}$$

according to Lemma 8. Letting $s > 1$ be an arbitrary constant, we have that

$$\begin{aligned} &\mathbb{P}\left\{\frac{\|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{C}^T\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F} \leq 1 + s\epsilon\right\} \\ &\geq \mathbb{P}\left\{\frac{\|\mathbf{A} - \mathbf{C}\mathbf{U}\mathbf{C}^T\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F} \leq 1 + s\epsilon \mid \frac{\|\mathbf{A} - \mathcal{P}_{\mathbf{C}_1}\mathbf{A}\|_F^2}{\|\mathbf{A} - \mathbf{A}_k\|_F^2} \leq f\right\} \\ &\quad \cdot \mathbb{P}\left\{\frac{\|\mathbf{A} - \mathcal{P}_{\mathbf{C}_1}\mathbf{A}\|_F^2}{\|\mathbf{A} - \mathbf{A}_k\|_F^2} \leq f\right\} \\ &\geq \left(1 - \frac{1 + \epsilon}{1 + s\epsilon}\right)(1 - 2\delta). \end{aligned}$$

where the last inequality follows from Lemma 7 and Lemma 8.

Repeating the sampling procedure for t times and letting $\mathbf{C}_{[i]}$ and $\mathbf{U}_{[i]}$ be the i -th sample, we obtain an upper error bound on the failure probability:

$$\begin{aligned} &\mathbb{P}\left\{\min_{i \in [t]} \left\{\frac{\|\mathbf{A} - \mathbf{C}_{[i]}\mathbf{U}_{[i]}\mathbf{C}_{[i]}^T\|_F}{\|\mathbf{A} - \mathbf{A}_k\|_F}\right\} \geq 1 + s\epsilon\right\} \\ &\leq \left(1 - \left(1 - \frac{1 + \epsilon}{1 + s\epsilon}\right)(1 - 2\delta)\right)^t \\ &= \left(1 + \frac{(s-1)(1-2\delta)}{\epsilon^{-1} + 1 + 2\delta(s-1)}\right)^{-t} \triangleq p. \end{aligned}$$

Taking logarithm of both sides of the equality and applying $\log(1+x) \approx x$ when x is small, we have

$$\begin{aligned} t &= \left\lceil \log\left(1 + \frac{(1-2\delta)(s-1)}{\epsilon^{-1} + 1 + 2\delta(s-1)}\right) \right\rceil^{-1} \log \frac{1}{p} \\ &\approx \frac{\epsilon^{-1} + 1 + 2\delta(s-1)}{(1-2\delta)(s-1)} \log \frac{1}{p}. \end{aligned}$$

Setting $s = 2$, we have that $t \approx (10\epsilon^{-1} + 18) \log(1/p)$.

Hence by sampling totally

$$c = (1 + \epsilon^{-1})(5k\epsilon^{-1} + 8.7\mu_k k \log(\sqrt{5}k))$$

columns and repeating the procedure for

$$t \geq (10\epsilon^{-1} + 18) \log(1/p)$$

times, the algorithm attains the upper error bound

$$\|\mathbf{A} - \mathbf{C}(\mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T) \mathbf{C}^T\|_F \leq (1 + 2\epsilon)\|\mathbf{A} - \mathbf{A}_k\|_F$$

with probability at least $1 - p$. Substituting 2ϵ by ϵ' yields the error bound in the theorem.

Time complexity and space complexity of Algorithm 1 is calculated as follows. The uniform sampling costs $\mathcal{O}(m)$ time; the first adaptive sampling round costs $\mathcal{O}(mc_1^2) + T_{\text{Multiply}}(m^2 c_1)$ time; the second adaptive sampling round costs $\mathcal{O}(m(c_1 + c_2)^2) + T_{\text{Multiply}}(m^2(c_1 + c_2))$ time; computing the intersection matrix costs $\mathcal{O}(mc^2) +$

$T_{\text{Multiply}}(m^2c)$ time in general. So the total time complexity is $\mathcal{O}(mc^2) + T_{\text{Multiply}}(m^2c)$ without using Theorem 4, or $\mathcal{O}(m(c_1 + c_2)^2) + T_{\text{Multiply}}(m^2c)$ using Theorem 4. As for the space complexity, the Moore-Penrose inverse of an $m \times c$ matrix demands $\mathcal{O}(mc)$ space, and multiplying a $c \times m$ matrix \mathbf{C}^\dagger by an $m \times m$ matrix \mathbf{A} costs $\mathcal{O}(mc)$ space by partition \mathbf{A} into small blocks of size smaller than $m \times c$ and loading one block into RAM at a time to perform matrix multiplication. \square

B Proof of Theorem 4

Proof. Let $\mathbf{C} \in \mathbb{R}^{m \times c}$ consists of a subset of columns of \mathbf{A} . By row permutation \mathbf{C} can be expressed as

$$\mathbf{PC} = \begin{bmatrix} \mathbf{W} \\ \mathbf{A}_{21} \end{bmatrix}.$$

Then according to Lemma 10, the Moore-Penrose inverse of \mathbf{C} can be written as

$$\mathbf{C}^\dagger = \mathbf{W}^{-1}(\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix} \mathbf{P},$$

where $\mathbf{S} = \mathbf{A}_{21} \mathbf{W}^{-1}$. Then the intersection matrix of modified Nyström approximation to \mathbf{A} can be expressed as

$$\begin{aligned} \mathbf{U} &= \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T \\ &= \mathbf{W}^{-1}(\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix} \mathbf{P} \mathbf{A} \mathbf{P}^T \\ &\quad \begin{bmatrix} \mathbf{I}_c \\ \mathbf{S} \end{bmatrix} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \mathbf{W}^{-1} \\ &= \mathbf{W}^{-1}(\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix} \\ &\quad \begin{bmatrix} \mathbf{W} & \mathbf{A}_{21}^T \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_c \\ \mathbf{S} \end{bmatrix} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \mathbf{W}^{-1} \\ &= \mathbf{W}^{-1}(\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \left(\mathbf{W} + \mathbf{A}_{21}^T \mathbf{S} + (\mathbf{A}_{21}^T \mathbf{S})^T \right. \\ &\quad \left. + \mathbf{S}^T \mathbf{A}_{22} \mathbf{S} \right) (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \mathbf{W}^{-1} \\ &\triangleq \mathbf{T}_1 (\mathbf{W} + \mathbf{T}_2 + \mathbf{T}_2^T + \mathbf{T}_3) \mathbf{T}_1^T. \end{aligned}$$

Here the intermediate matrices are computed by

$$\begin{aligned} \mathbf{T}_0 &= \mathbf{A}_{21}^T \mathbf{A}_{21}, \\ \mathbf{T}_1 &= \mathbf{W}^{-1}(\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \\ &= \mathbf{W}^{-1} \left(\mathbf{I}_c + \mathbf{W}^{-1} \mathbf{T}_0 \mathbf{W}^{-1} \right)^{-1}, \\ \mathbf{T}_2 &= \mathbf{A}_{21}^T \mathbf{S} = \mathbf{A}_{21}^T \mathbf{A}_{21} \mathbf{W}^{-1} = \mathbf{T}_0 \mathbf{W}^{-1}, \\ \mathbf{T}_3 &= \mathbf{S}^T \mathbf{A}_{22} \mathbf{S} = \mathbf{W}^{-1} \left(\mathbf{A}_{21}^T \mathbf{A}_{22} \mathbf{A}_{21} \right) \mathbf{W}^{-1}. \end{aligned}$$

The matrix inverse operations are on $c \times c$ matrices which costs $\mathcal{O}(c^3)$ time. The matrix multiplication $\mathbf{A}_{21}^T \mathbf{A}_{22} \mathbf{A}_{21}$ requires time $T_{\text{Multiply}}((m - c)^2 c)$. \square

Lemma 10 (The Moore Penrose Inverse of Partitioned Matrices (Ben-Israel and Greville, 2003, Page 179)). *Given a matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ of rank of at least c which has a nonsingular $c \times c$ submatrix \mathbf{X}_{11} . By rearrangement of columns and rows by permutation matrices \mathbf{P} and \mathbf{Q} , the submatrix \mathbf{X}_{11} can be brought to the top left corner of \mathbf{X} , that is,*

$$\mathbf{P} \mathbf{X} \mathbf{Q} = \begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{21} & \mathbf{X}_{22} \end{bmatrix}.$$

Then the Moore-Penrose inverse of \mathbf{X} is

$$\mathbf{X}^\dagger = \mathbf{Q} \begin{bmatrix} \mathbf{I}_c \\ \mathbf{T}^T \end{bmatrix} (\mathbf{I}_c + \mathbf{T} \mathbf{T}^T)^{-1} \mathbf{X}_{11}^{-1} (\mathbf{I}_c + \mathbf{S} \mathbf{S}^T)^{-1} \begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix} \mathbf{P},$$

where $\mathbf{T} = \mathbf{X}_{11}^{-1} \mathbf{X}_{12}$ and $\mathbf{S} = \mathbf{X}_{21} \mathbf{X}_{11}^{-1}$.

C The Proof of Theorem 5

Proof. Suppose that $\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{A})$. We have that $\text{rank}(\mathbf{W}) = \text{rank}(\mathbf{C}) = \text{rank}(\mathbf{A})$ because

$$\text{rank}(\mathbf{A}) \geq \text{rank}(\mathbf{C}) \geq \text{rank}(\mathbf{W}). \quad (9)$$

Thus there exists a matrix \mathbf{X} such that

$$\begin{bmatrix} \mathbf{A}_{21}^T \\ \mathbf{A}_{22} \end{bmatrix} = \mathbf{C} \mathbf{X}^T = \begin{bmatrix} \mathbf{W} \mathbf{X}^T \\ \mathbf{A}_{21} \mathbf{X}^T \end{bmatrix},$$

and it follows that $\mathbf{A}_{21} = \mathbf{X} \mathbf{W}$ and $\mathbf{A}_{22} = \mathbf{A}_{21} \mathbf{X}^T = \mathbf{X} \mathbf{W} \mathbf{X}^T$. Then we have that

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \mathbf{W} & (\mathbf{X} \mathbf{W})^T \\ \mathbf{X} \mathbf{W} & \mathbf{X} \mathbf{W} \mathbf{X}^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} \\ \mathbf{X} \end{bmatrix} \mathbf{W} \begin{bmatrix} \mathbf{I} & \mathbf{X}^T \end{bmatrix}, \quad (10) \\ \mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T &= \begin{bmatrix} \mathbf{W} \\ \mathbf{X} \mathbf{W} \end{bmatrix} \mathbf{W}^\dagger \begin{bmatrix} \mathbf{W} & (\mathbf{X} \mathbf{W})^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} \\ \mathbf{X} \end{bmatrix} \mathbf{W} \begin{bmatrix} \mathbf{I} & \mathbf{X}^T \end{bmatrix}. \quad (11) \end{aligned}$$

Here the second equality in (11) follows from $\mathbf{W} \mathbf{W}^\dagger \mathbf{W} = \mathbf{W}$. We obtain that $\mathbf{A} = \mathbf{C} \mathbf{W}^\dagger \mathbf{C}$. Then we show that $\mathbf{A} = \mathbf{C} \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T \mathbf{C}^T$.

Since $\mathbf{C}^\dagger = (\mathbf{C}^T \mathbf{C})^\dagger \mathbf{C}^T$, we have that

$$\mathbf{C}^\dagger = (\mathbf{W}(\mathbf{I} + \mathbf{X}^T \mathbf{X}) \mathbf{W})^\dagger \mathbf{W} \begin{bmatrix} \mathbf{I} & \mathbf{X}^T \end{bmatrix},$$

and thus

$$\begin{aligned} \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T \mathbf{W} &= (\mathbf{W}(\mathbf{I} + \mathbf{X}^T \mathbf{X}) \mathbf{W})^\dagger \mathbf{W} (\mathbf{I} + \mathbf{X}^T \mathbf{X}) \begin{bmatrix} \mathbf{W}(\mathbf{I} + \mathbf{X}^T \mathbf{X}) \\ \mathbf{W}(\mathbf{W}(\mathbf{I} + \mathbf{X}^T \mathbf{X}) \mathbf{W})^\dagger \mathbf{W} \end{bmatrix} \\ &= (\mathbf{W}(\mathbf{I} + \mathbf{X}^T \mathbf{X}) \mathbf{W})^\dagger \mathbf{W} (\mathbf{I} + \mathbf{X}^T \mathbf{X}) \mathbf{W}, \end{aligned}$$

where the second equality follows from Lemma 11 because $(\mathbf{I} + \mathbf{X}^T \mathbf{X})$ is positive definite. Similarly we have

$$\begin{aligned} & \mathbf{W} \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T \mathbf{W} \\ &= \mathbf{W} (\mathbf{W} (\mathbf{I} + \mathbf{X}^T \mathbf{X}) \mathbf{W})^\dagger \mathbf{W} (\mathbf{I} + \mathbf{X}^T \mathbf{X}) \mathbf{W} = \mathbf{W}. \end{aligned}$$

Thus we have

$$\begin{aligned} \mathbf{C} \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T \mathbf{C} &= \begin{bmatrix} \mathbf{I} \\ \mathbf{X} \end{bmatrix} \mathbf{W} \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T \mathbf{W} \begin{bmatrix} \mathbf{I} & \mathbf{X}^T \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I} \\ \mathbf{X} \end{bmatrix} \mathbf{W} \begin{bmatrix} \mathbf{I} & \mathbf{X}^T \end{bmatrix}. \end{aligned} \quad (12)$$

It follows from Equations (10) (11) (12) that $\mathbf{A} = \mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T = \mathbf{C} \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T \mathbf{C}^T$.

Conversely, when $\mathbf{A} = \mathbf{C} \mathbf{W}^\dagger \mathbf{C}^T$, we have that $\text{rank}(\mathbf{A}) \leq \text{rank}(\mathbf{W}^\dagger) = \text{rank}(\mathbf{W})$. By applying (9) we have that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{W})$.

When $\mathbf{A} = \mathbf{C} \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^\dagger)^T \mathbf{C}^T$, we have $\text{rank}(\mathbf{A}) \leq \text{rank}(\mathbf{C})$. Thus there exists a matrix \mathbf{X} such that

$$\begin{bmatrix} \mathbf{A}_{21}^T \\ \mathbf{A}_{22} \end{bmatrix} = \mathbf{C} \mathbf{X}^T = \begin{bmatrix} \mathbf{W} \mathbf{X}^T \\ \mathbf{A}_{21} \mathbf{X}^T \end{bmatrix},$$

and therefore $\mathbf{A}_{21} = \mathbf{X} \mathbf{W}$. Then we have that

$$\mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{A}_{21} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \mathbf{X} \end{bmatrix} \mathbf{W},$$

so $\text{rank}(\mathbf{C}) \leq \text{rank}(\mathbf{W})$. Apply (9) again we have $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{W})$. \square

Lemma 11. $\mathbf{X}^T \mathbf{V} \mathbf{X} (\mathbf{X}^T \mathbf{V} \mathbf{X})^\dagger \mathbf{X}^T = \mathbf{X}^T$ for any positive definite matrix \mathbf{V} .

Proof. Since the positive definite matrix \mathbf{V} have a decomposition $\mathbf{V} = \mathbf{B}^T \mathbf{B}$ for some nonsingular matrix \mathbf{B} , so we have

$$\begin{aligned} & \mathbf{X}^T \mathbf{V} \mathbf{X} (\mathbf{X}^T \mathbf{V} \mathbf{X})^\dagger \mathbf{X}^T \\ &= (\mathbf{B} \mathbf{X})^T \left(\mathbf{B} \mathbf{X} ((\mathbf{B} \mathbf{X})^T (\mathbf{B} \mathbf{X}))^\dagger \right) (\mathbf{B} \mathbf{X})^T \mathbf{B} (\mathbf{B}^T \mathbf{B})^{-1} \\ &= (\mathbf{B} \mathbf{X})^T ((\mathbf{B} \mathbf{X})^T)^\dagger (\mathbf{B} \mathbf{X})^T (\mathbf{B}^T)^{-1} \\ &= (\mathbf{B} \mathbf{X})^T (\mathbf{B}^T)^{-1} \\ &= \mathbf{X}^T. \end{aligned}$$

\square

D Proof of Theorem 6

In Section D.1 we provide two key lemmas, and then in Section D.2 we prove Theorem 6 using the two lemmas.

D.1 Key Lemmas

Lemma 12. For an $m \times m$ matrix \mathbf{B} with diagonal entries equal to one and off-diagonal entries equal to α , the error incurred by the modified Nystrom method is lower bounded by

$$\begin{aligned} & \|\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{mod}}\|_F^2 \\ & \geq (1 - \alpha)^2 (m - c) \left(1 + \frac{2}{c} - (1 - \alpha) \frac{1 + o(1)}{\alpha c m / 2} \right). \end{aligned}$$

Proof. Without loss of generality, we assume the first c column of \mathbf{B} are selected to construct \mathbf{C} . We partition \mathbf{B} and \mathbf{C} as:

$$\mathbf{B} = \begin{bmatrix} \mathbf{W} & \mathbf{B}_{21}^T \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{B}_{21} \end{bmatrix}.$$

Here the matrix \mathbf{W} can be expressed by $\mathbf{W} = (1 - \alpha) \mathbf{I}_c + \alpha \mathbf{1}_c \mathbf{1}_c^T$. We apply the Sherman-Morrison-Woodbury formula

$$(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{B} (\mathbf{C}^{-1} + \mathbf{DA}^{-1} \mathbf{B})^{-1} \mathbf{DA}^{-1}$$

to compute \mathbf{W}^{-1} , yielding

$$\mathbf{W}^{-1} = \frac{1}{1 - \alpha} \mathbf{I}_c - \frac{\alpha}{(1 - \alpha)(1 - \alpha + c\alpha)} \mathbf{1}_c \mathbf{1}_c^T. \quad (13)$$

We expand the Moore-Penrose inverse of \mathbf{C} by Lemma 10 and obtain

$$\mathbf{C}^\dagger = \mathbf{W}^{-1} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix}$$

where

$$\mathbf{S} = \mathbf{B}_{21} \mathbf{W}^{-1} = \frac{\alpha}{1 - \alpha + c\alpha} \mathbf{1}_{m-c} \mathbf{1}_c^T.$$

It is easily verified that $\mathbf{S}^T \mathbf{S} = \left(\frac{\alpha}{1 - \alpha + c\alpha} \right)^2 (m - c) \mathbf{1}_c \mathbf{1}_c^T$.

Now we express the matrix constructed by the modified Nystrom method in a partitioned form:

$$\begin{aligned} \tilde{\mathbf{B}}_c^{\text{mod}} &= \mathbf{C} \mathbf{C}^\dagger \mathbf{B} (\mathbf{C}^\dagger)^T \mathbf{C}^T \\ &= \begin{bmatrix} \mathbf{W} \\ \mathbf{B}_{21} \end{bmatrix} \mathbf{W}^{-1} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix} \mathbf{B} \\ &= \begin{bmatrix} \mathbf{I}_c \\ \mathbf{S} \end{bmatrix} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \mathbf{W}^{-1} \begin{bmatrix} \mathbf{W} \\ \mathbf{B}_{21} \end{bmatrix}^T \\ &= \begin{bmatrix} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \\ \mathbf{B}_{21} \mathbf{W}^{-1} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix} \mathbf{B} \\ &= \begin{bmatrix} \mathbf{I}_c \\ \mathbf{S} \end{bmatrix} \begin{bmatrix} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \\ \mathbf{B}_{21} \mathbf{W}^{-1} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} \end{bmatrix}^T. \end{aligned} \quad (14)$$

We then compute the submatrices $(\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1}$ and $\mathbf{B}_{21} \mathbf{W}^{-1} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1}$ respectively as follows. We apply

the Sherman-Morrison-Woodbury formula to compute $(\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1}$, yielding

$$\begin{aligned} (\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} &= \left(\mathbf{I}_c + \left(\frac{\alpha}{1 - \alpha + c\alpha} \right)^2 (m - c) \mathbf{1}_c \mathbf{1}_c^T \right)^{-1} \\ &= \mathbf{I}_c - \gamma_1 \mathbf{1}_c \mathbf{1}_c^T, \end{aligned} \quad (15)$$

where

$$\gamma_1 = \frac{m - c}{mc + \left(\frac{1 - \alpha}{\alpha} \right)^2 + \frac{2(1 - \alpha)c}{\alpha}}.$$

It follows from (13) and (15) that

$$\begin{aligned} \mathbf{W}^{-1}(\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} &= (\gamma_2 \mathbf{I}_c - \gamma_3 \mathbf{1}_c \mathbf{1}_c^T)(\mathbf{I}_c - \gamma_1 \mathbf{1}_c \mathbf{1}_c^T) \\ &= \gamma_2 \mathbf{I}_c + (\gamma_1 \gamma_3 c - \gamma_1 \gamma_2 - \gamma_3) \mathbf{1}_c \mathbf{1}_c^T \end{aligned} \quad (16)$$

where

$$\gamma_2 = \frac{1}{1 - \alpha} \quad \text{and} \quad \gamma_3 = \frac{\alpha}{(1 - \alpha)(1 - \alpha + c\alpha)}.$$

Then we have that

$$\begin{aligned} \mathbf{B}_{21} \mathbf{W}^{-1}(\mathbf{I}_c + \mathbf{S}^T \mathbf{S})^{-1} &= \alpha(\gamma_1 \gamma_3 c^2 - \gamma_3 c - \gamma_1 \gamma_2 c + \gamma_2) \mathbf{1}_{m-c} \mathbf{1}_c^T \\ &\triangleq \gamma \mathbf{1}_{m-c} \mathbf{1}_c^T, \end{aligned} \quad (17)$$

where

$$\begin{aligned} \gamma &= \alpha(\gamma_1 \gamma_3 c^2 - \gamma_3 c - \gamma_1 \gamma_2 c + \gamma_2) \\ &= \frac{\alpha(\alpha c - \alpha + 1)}{2\alpha c - 2\alpha - 2\alpha^2 c + \alpha^2 + \alpha^2 c m + 1}. \end{aligned} \quad (18)$$

Since $\mathbf{B}_{21} = \alpha \mathbf{1}_{m-c} \mathbf{1}_c^T$ and $\mathbf{B}_{22} = (1 - \alpha) \mathbf{I}_{m-c} + \alpha \mathbf{1}_{m-c} \mathbf{1}_{m-c}^T$, it is easily verified that

$$\begin{aligned} \begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix} \mathbf{B} \begin{bmatrix} \mathbf{I}_c \\ \mathbf{S} \end{bmatrix} &= \begin{bmatrix} \mathbf{I}_c & \mathbf{S}^T \end{bmatrix} \begin{bmatrix} \mathbf{W} & \mathbf{B}_{21}^T \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_c \\ \mathbf{S} \end{bmatrix} \\ &= (1 - \alpha) \mathbf{I}_c + \lambda \mathbf{1}_c \mathbf{1}_c^T, \end{aligned} \quad (19)$$

where

$$\lambda = \frac{\alpha(3\alpha m - \alpha c - 2\alpha + \alpha^2 c - 3\alpha^2 m + \alpha^2 + \alpha^2 m^2 + 1)}{(\alpha c - \alpha + 1)^2}$$

It follows from (14), (15), (17), and (19) that

$$\begin{aligned} \tilde{\mathbf{B}}_c^{\text{mod}} &= \begin{bmatrix} \mathbf{I}_c - \gamma_1 \mathbf{1}_c \mathbf{1}_c^T \\ \gamma \mathbf{1}_{m-c} \mathbf{1}_c^T \end{bmatrix} \left((1 - \alpha) \mathbf{I}_c + \lambda \mathbf{1}_c \mathbf{1}_c^T \right) \begin{bmatrix} \mathbf{I}_c - \gamma_1 \mathbf{1}_c \mathbf{1}_c^T \\ \gamma \mathbf{1}_{m-c} \mathbf{1}_c^T \end{bmatrix}^T \\ &\triangleq \begin{bmatrix} \tilde{\mathbf{B}}_{11} & \tilde{\mathbf{B}}_{21}^T \\ \tilde{\mathbf{B}}_{21} & \tilde{\mathbf{B}}_{22} \end{bmatrix}, \end{aligned}$$

where

$$\begin{aligned} \tilde{\mathbf{B}}_{11} &= (1 - \alpha) \mathbf{I}_c + [(1 - \gamma_1 c) \\ &\quad (\lambda - \lambda \gamma_1 c - (1 - \alpha) \gamma_1) - (1 - \alpha) \gamma_1] \mathbf{1}_c \mathbf{1}_c^T \\ &= (1 - \alpha) \mathbf{I}_c + \eta_1 \mathbf{1}_c \mathbf{1}_c^T, \\ \tilde{\mathbf{B}}_{21} &= \tilde{\mathbf{A}}_{12}^T = \gamma(1 - \gamma_1 c)(1 - \alpha + \lambda c) \mathbf{1}_{m-c} \mathbf{1}_c^T \\ &= \eta_2 \mathbf{1}_{m-c} \mathbf{1}_c^T, \\ \tilde{\mathbf{B}}_{22} &= \gamma^2 c(1 - \alpha + \lambda c) \mathbf{1}_{m-c} \mathbf{1}_{m-c}^T \\ &= \eta_3 \mathbf{1}_{m-c} \mathbf{1}_{m-c}^T, \end{aligned}$$

where

$$\begin{aligned} \eta_1 &= (1 - \gamma_1 c)(\lambda - \lambda \gamma_1 c - (1 - \alpha) \gamma_1) - (1 - \alpha) \gamma_1, \\ \eta_2 &= \gamma(1 - \gamma_1 c)(1 - \alpha + \lambda c), \\ \eta_3 &= \gamma^2 c(1 - \alpha + \lambda c), \end{aligned}$$

By dealing with the four blocks of $\tilde{\mathbf{B}}_c^{\text{mod}}$ respectively, we finally obtain that

$$\begin{aligned} \|\mathbf{B} - \tilde{\mathbf{B}}_c^{\text{mod}}\|_F^2 &= \|\mathbf{W} - \tilde{\mathbf{B}}_{11}\|_F^2 + 2\|\mathbf{B}_{21} - \tilde{\mathbf{B}}_{21}\|_F^2 + \|\mathbf{B}_{22} - \tilde{\mathbf{B}}_{22}\|_F^2 \\ &= c^2(\alpha - \eta_1)^2 + 2c(m - c)(\alpha - \eta_2)^2 \\ &\quad + (m - c)(m - c - 1)(\alpha - \eta_3)^2 + (m - c)(1 - \eta_3)^2 \\ &= (m - c)(\alpha - 1)^2(\alpha^4 c^2 m^2 - 4\alpha^4 c^2 m + 4\alpha^4 c^2 \\ &\quad + 2\alpha^4 c m^2 - 4\alpha^4 c m + \alpha^4 c + \alpha^4 m - \alpha^4 + 4\alpha^3 c^2 m \\ &\quad - 8\alpha^3 c^2 + 2\alpha^3 c m + 2\alpha^3 c - 2\alpha^3 m + 2\alpha^3 + 4\alpha^2 c^2 \\ &\quad + 2\alpha^2 c m - 7\alpha^2 c + \alpha^2 m + 4\alpha c - 2\alpha + 1)/(2\alpha c \\ &\quad - 2\alpha - 2\alpha^2 c + \alpha^2 + \alpha^2 c m + 1)^2 \\ &= (m - c)(\alpha - 1)^2 \left(1 + \frac{2}{c} - \frac{(1 - \alpha)}{c} (6\alpha c - 6\alpha \right. \\ &\quad - 12\alpha^2 c + 6\alpha^3 c + 6\alpha^2 - 2\alpha^3 + 3\alpha^2 c^2 - 3\alpha^3 c^2 \\ &\quad + 2\alpha^3 c^2 m + 3\alpha^2 c m - 3\alpha^3 c m + 2)/(2\alpha c - 2\alpha \\ &\quad \left. - 2\alpha^2 c + \alpha^2 + \alpha^2 c m + 1)^2 \right) \\ &= (m - c)(\alpha - 1)^2 \left(1 + \frac{2}{c} - (1 + o(1)) \frac{1 - \alpha}{\alpha c m / 2} \right). \end{aligned}$$

□

Lemma 13 (Lemma 19 of Wang and Zhang (2013)). *Given m and k , we let \mathbf{B} be an $\frac{m}{k} \times \frac{m}{k}$ matrix whose diagonal entries equal to one and off-diagonal entries equal to $\alpha \in [0, 1)$. We let \mathbf{A} be an $m \times m$ block-diagonal matrix*

$$\mathbf{A} = \text{diag}(\underbrace{\mathbf{B}, \dots, \mathbf{B}}_{k \text{ blocks}}). \quad (20)$$

Let \mathbf{A}_k be the best rank- k approximation to the matrix \mathbf{A} , then we have that

$$\|\mathbf{A} - \mathbf{A}_k\|_F = (1 - \alpha) \sqrt{m - k}.$$

D.2 Proof of the Theorem

Now we prove Theorem 6 using Lemma 12 and Lemma 13.

Proof. Let \mathbf{C} consist of c column sampled from \mathbf{A} and $\hat{\mathbf{C}}_i$ consist of c_i columns sampled from the i -th block diagonal matrix in \mathbf{A} . Without loss of generality, we assume $\hat{\mathbf{C}}_i$ consists of the first c_i columns of \mathbf{B} . Then the intersection matrix \mathbf{U} is computed by

$$\begin{aligned} \mathbf{U} &= \mathbf{C}^\dagger \mathbf{A} (\mathbf{C}^T)^\dagger \\ &= [\text{diag}(\hat{\mathbf{C}}_1, \dots, \hat{\mathbf{C}}_k)]^\dagger \mathbf{A} [\text{diag}(\hat{\mathbf{C}}_1^T, \dots, \hat{\mathbf{C}}_k^T)]^\dagger \\ &= \text{diag}(\hat{\mathbf{C}}_1^\dagger \mathbf{B} (\hat{\mathbf{C}}_1^\dagger)^T, \dots, \hat{\mathbf{C}}_k^\dagger \mathbf{B} (\hat{\mathbf{C}}_k^\dagger)^T). \end{aligned}$$

The modified Nyström approximation to \mathbf{A} is

$$\begin{aligned} \tilde{\mathbf{A}}_c^{\text{mod}} &= \mathbf{C} \mathbf{U} \mathbf{C}^T \\ &= \text{diag}(\hat{\mathbf{C}}_1 \hat{\mathbf{C}}_1^\dagger \mathbf{B} (\hat{\mathbf{C}}_1^\dagger)^T \hat{\mathbf{C}}_1^T, \dots, \hat{\mathbf{C}}_k \hat{\mathbf{C}}_k^\dagger \mathbf{B} (\hat{\mathbf{C}}_k^\dagger)^T \hat{\mathbf{C}}_k^T), \end{aligned}$$

and thus the approximation error is

$$\begin{aligned} \|\mathbf{A} - \tilde{\mathbf{A}}_c^{\text{mod}}\|_F^2 &= \sum_{i=1}^k \left\| \mathbf{B} - \hat{\mathbf{C}}_i \hat{\mathbf{C}}_i^\dagger \mathbf{B} (\hat{\mathbf{C}}_i^\dagger)^T \hat{\mathbf{C}}_i^T \right\|_F^2 \\ &\geq (1 - \alpha)^2 \sum_{i=1}^k (p - c_i) \left(1 + \frac{2}{c_i} - (1 - \alpha) \left(\frac{1 + o(1)}{\alpha c_i p / 2} \right) \right) \\ &= (1 - \alpha)^2 \left(\sum_{i=1}^k (p - c_i) \right. \\ &\quad \left. + \sum_{i=1}^k \frac{2(p - c_i)}{c_i} \left(1 - \frac{(1 - \alpha)(1 + o(1))}{\alpha p} \right) \right) \\ &\geq (1 - \alpha)^2 (m - c) \left(1 + \frac{2k}{c} \left(1 - \frac{k(1 - \alpha)(1 + o(1))}{\alpha m} \right) \right), \end{aligned}$$

where the former inequality follows from Lemma 12, and the latter inequality follows by minimizing over c_1, \dots, c_k . Finally we apply Lemma 13, and the theorem follows by setting $\alpha \rightarrow 1$. \square

E Supplementary Experiments

We have mentioned in Remark 1 that the resulting approximation accuracy is insensitive to the parameter μ in Algorithm 1, and setting μ to be exactly the matrix coherence does not in general give rise to the highest accuracy. To demonstrate this point of view, we conduct experiments on an RBF kernel matrix of the Letters Dataset with $\sigma = 0.2$, and we set $k = 10$.

We compare the uniform+adaptive² algorithm with different settings of μ ; we also employ the adaptive-full algorithm of Kumar et al. (2012), the near-optimal+adaptive algorithm of Wang and Zhang (2013), and the uniform sampling algorithm for comparison. The experiment

settings are the same to Section 6. Here the adaptive-full algorithm also has three steps: one uniform sampling and two adaptive sampling steps, and we set $c_1 = c_2 = c_3 = c/3$ according to Kumar et al. (2012). We plot the approximation errors in Figure 5.

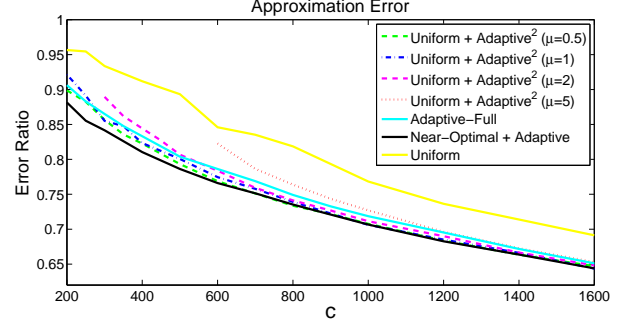


Figure 5: Effect of the parameter μ in Algorithm 1.

We can see from Figure 5 that different settings of μ does not have big influence on the approximation accuracy. We can also see that it is unnecessary to set μ to be exactly the matrix coherence; in this set of experiments, the uniform+adaptive² algorithm achieves the higher accuracy when $\mu = 0.5$ (the actual matrix coherence is $\mu_{10} = 62.05$).